
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM343

Minimizing the production impact of ML model updates with shadow testing

Tarun Sairam

Senior Product Manager,
Amazon SageMaker
AWS

Qingwei Li

Senior AI/ML Specialist Architect
AWS

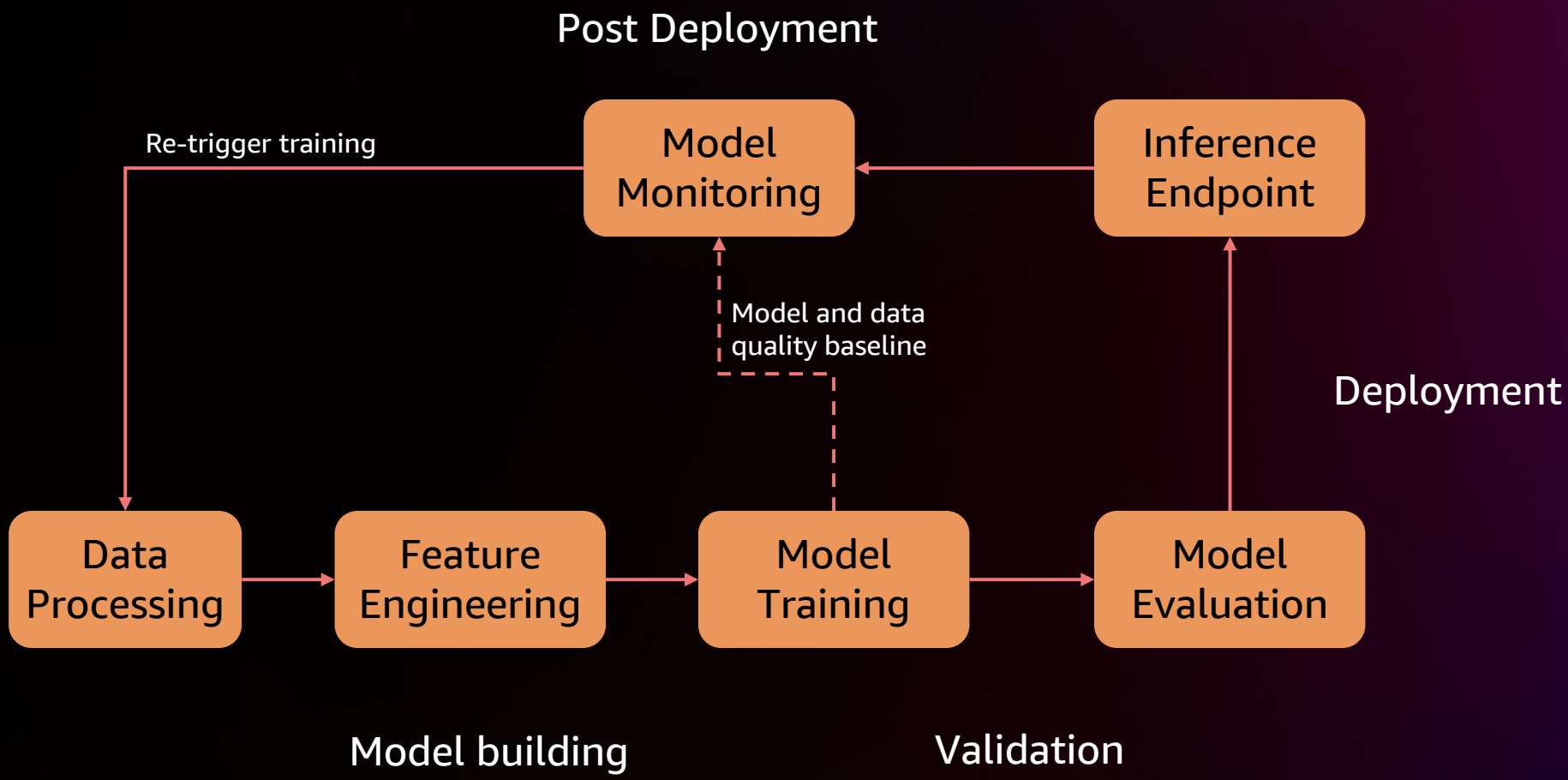
Vivek Govindapillai

Director, Software Engineering
HERE Technologies

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introduction: Typical ML lifecycle



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model validation and deployment options

	Training Validation	Offline Replays	Deployment Guardrails	Shadow testing
+	Readily available	Assess quality on production data	Auto rollbacks	Live production data without impact
-	Data drift	Operational metrics lost	Customer impact	Implementation complexity
$	$	$$	$	$$-$$$

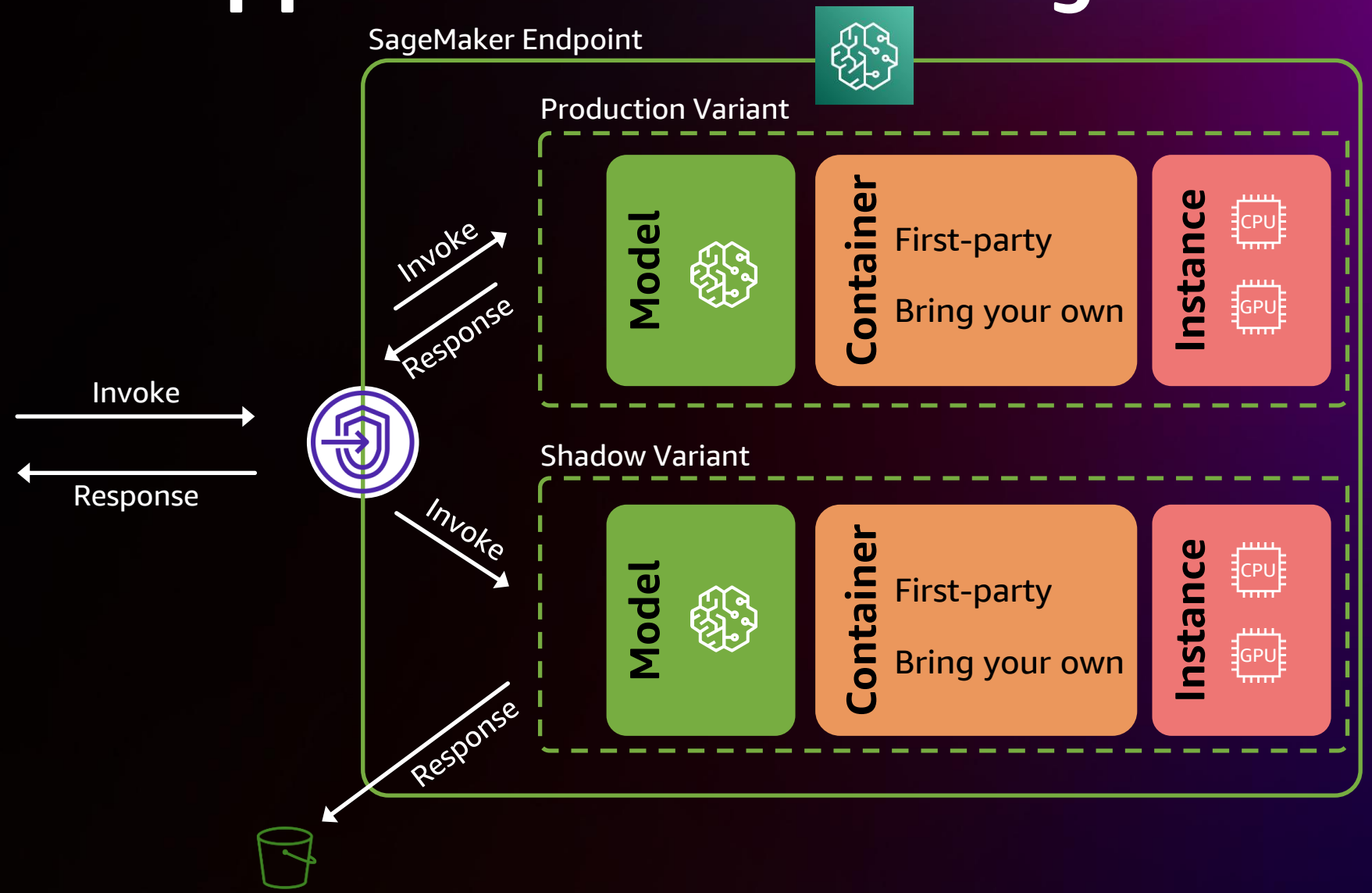
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker now supports shadow testing

SageMaker takes care of mirroring requests

Start small and dial up to control costs

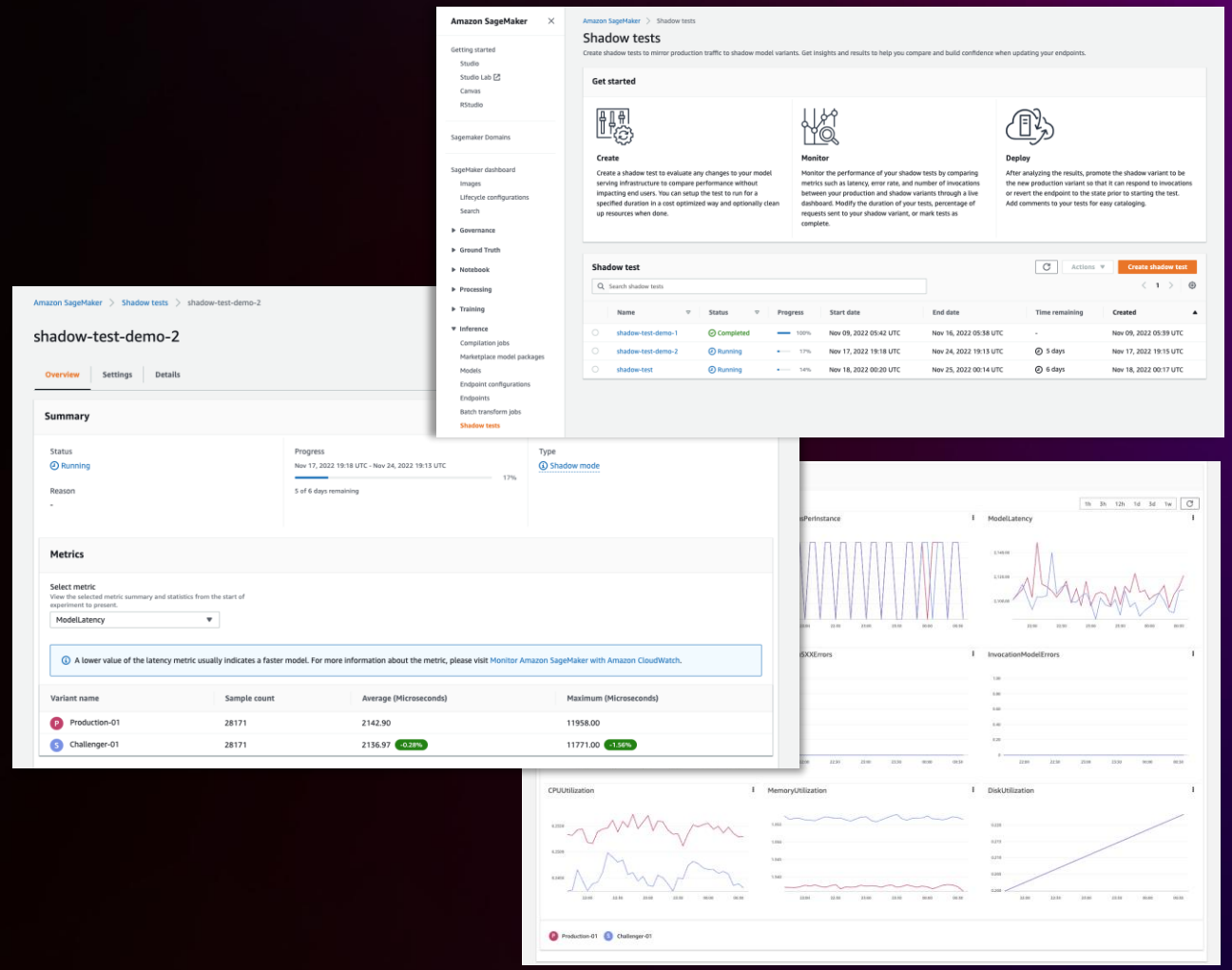
Accessible through AWS console, CLI, APIs



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Also introducing tools to manage shadow tests

- Set up a test for a predefined duration
- Monitor operational performance through a live dashboard
- Deploy models with confidence or rollback



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key benefits

Leverage existing workflows

Perform shadow testing without building custom orchestration



Accelerate model deployment

Reduce time to validate models from weeks to days

Optimize costs

Start small and dial up traffic as you gain confidence

Monitor

Compare operational performance through a live dashboard

Proactively validate

Validate changes to any component of your production variant

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Pricing and availability

SageMaker Inference now supports shadow testing



Pricing: Same as production variants; ML instances and storage



Availability: GA in Regions where SageMaker Inference is available (excluding AWS GovCloud (US) and China Regions)



Compatibility: Works with real-time inference endpoints

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker

Getting started

Control panel
- Studio
- Studio Lab
- Canvas
- RStudio

SageMaker dashboard
- Images
- Lifecycle configurations
- Search

Ground Truth

Notebook

Processing

Training

Inference

MACHINE LEARNING

Amazon SageMaker

Build, train, and deploy machine learning models at scale

The quickest and easiest way to get ML models from idea to production.

New to SageMaker?

Get started with Amazon SageMaker by completing the quick start guide.

Get Started

How it works

What is Amazon SageMaker?

Amazon SageMaker provides machine learning (ML) capabilities for data scientists and developers to prepare, build, train, and deploy high-quality ML models efficiently.

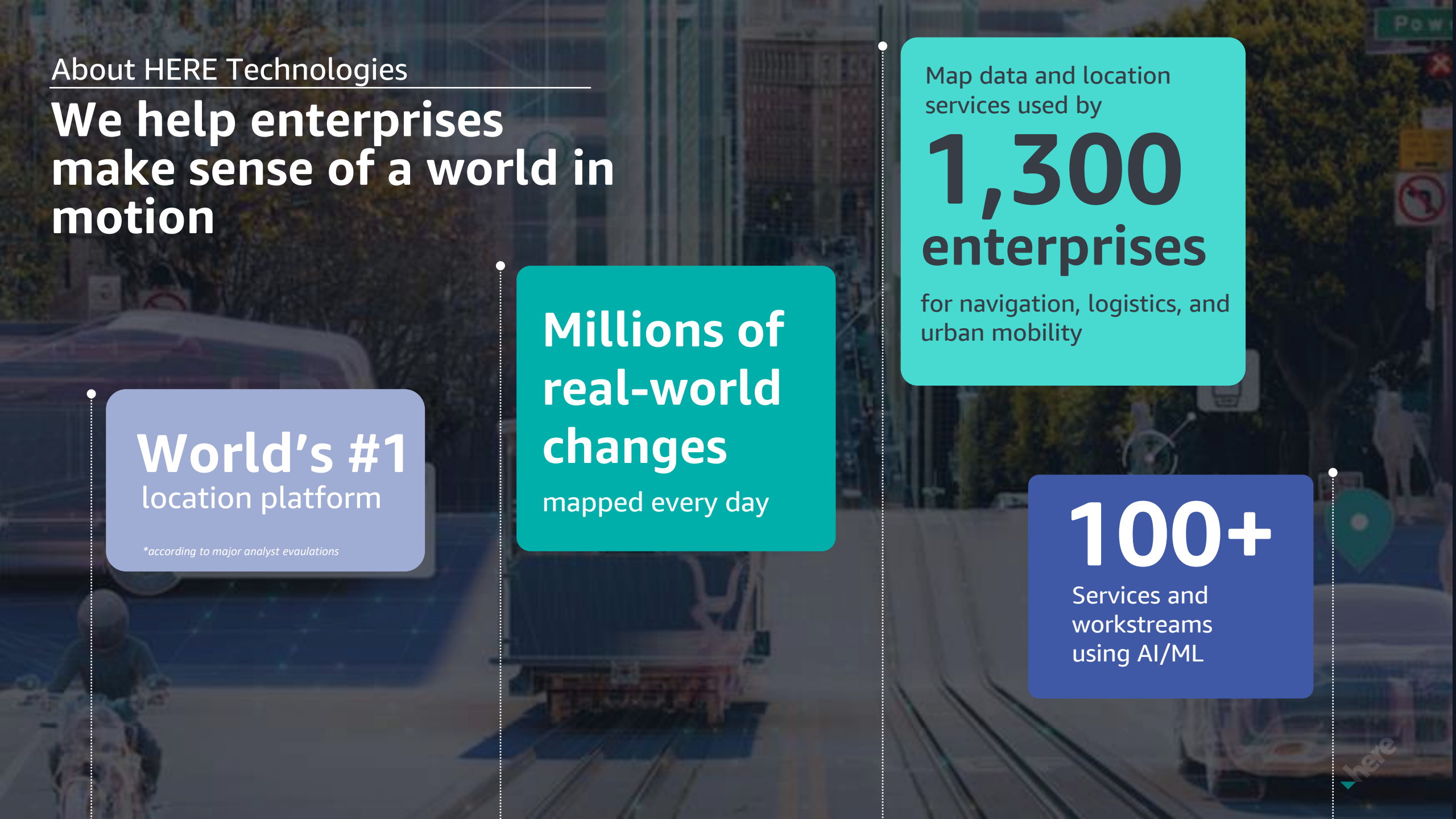
New user onboarding guide NEW

Documentation

- Getting started
- Tutorials
- Documentation
- Developer Resources
- AWS Developer Forum
- Contact us

HERE Technologies

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



About HERE Technologies

We help enterprises make sense of a world in motion

World's #1
location platform

**according to major analyst evaulations*

Millions of real-world changes
mapped every day

Map data and location services used by

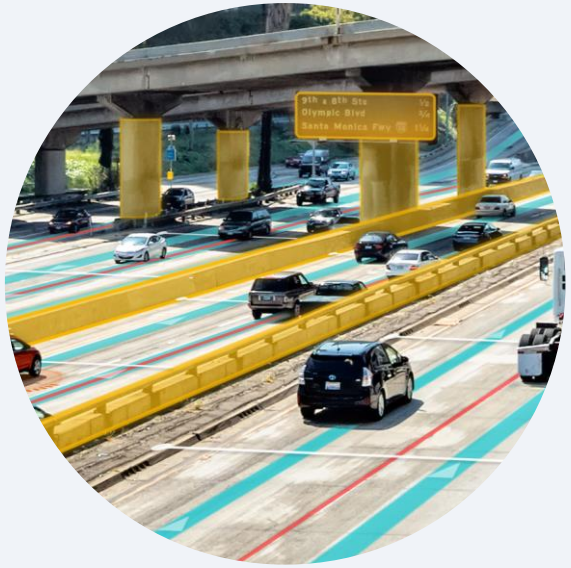
1,300
enterprises

for navigation, logistics, and urban mobility

100+
Services and workstreams using AI/ML

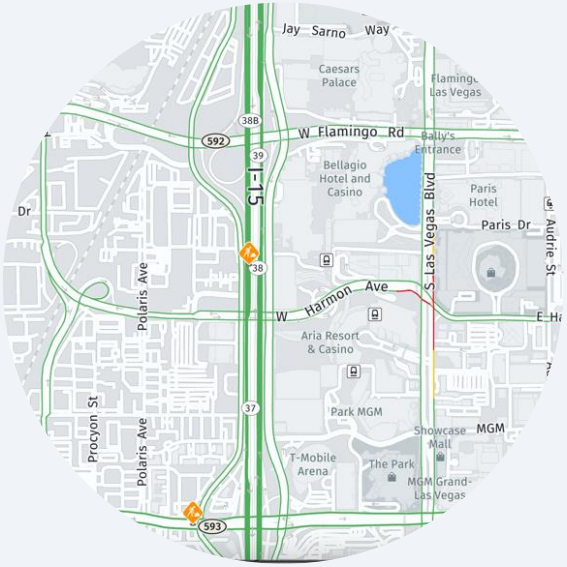
ML and location technology are key to understanding how the world might change in a minute, hour, or day from now

Aiding awareness for 170M+ vehicles



BMW 7 Series uses HERE high-definition map to enable hands-free driving in the U.S.

Predicting traffic and ETAs



USPS, FedEx, and Garmin use HERE to keep tabs on road conditions

Private maps for logistics



An ecommerce leader creates private map and routing layers to aid middle mile logistics



We used

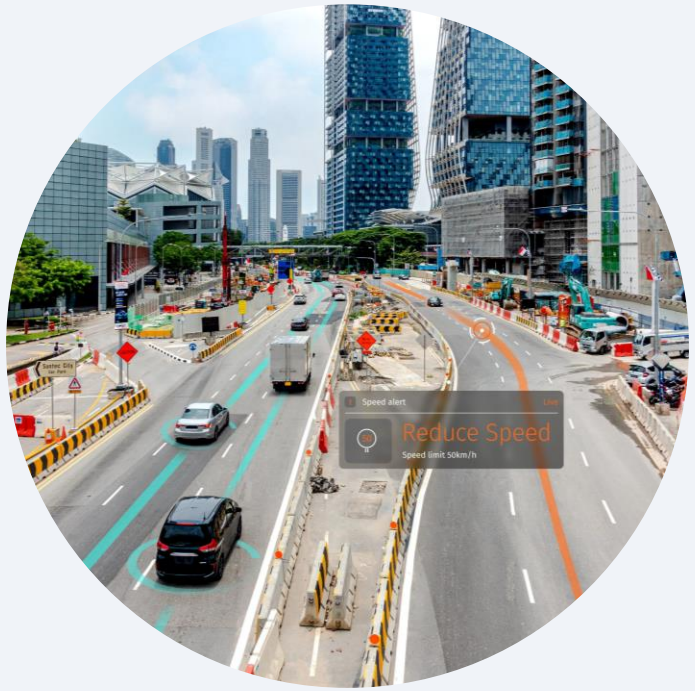
Amazon SageMaker's shadow test

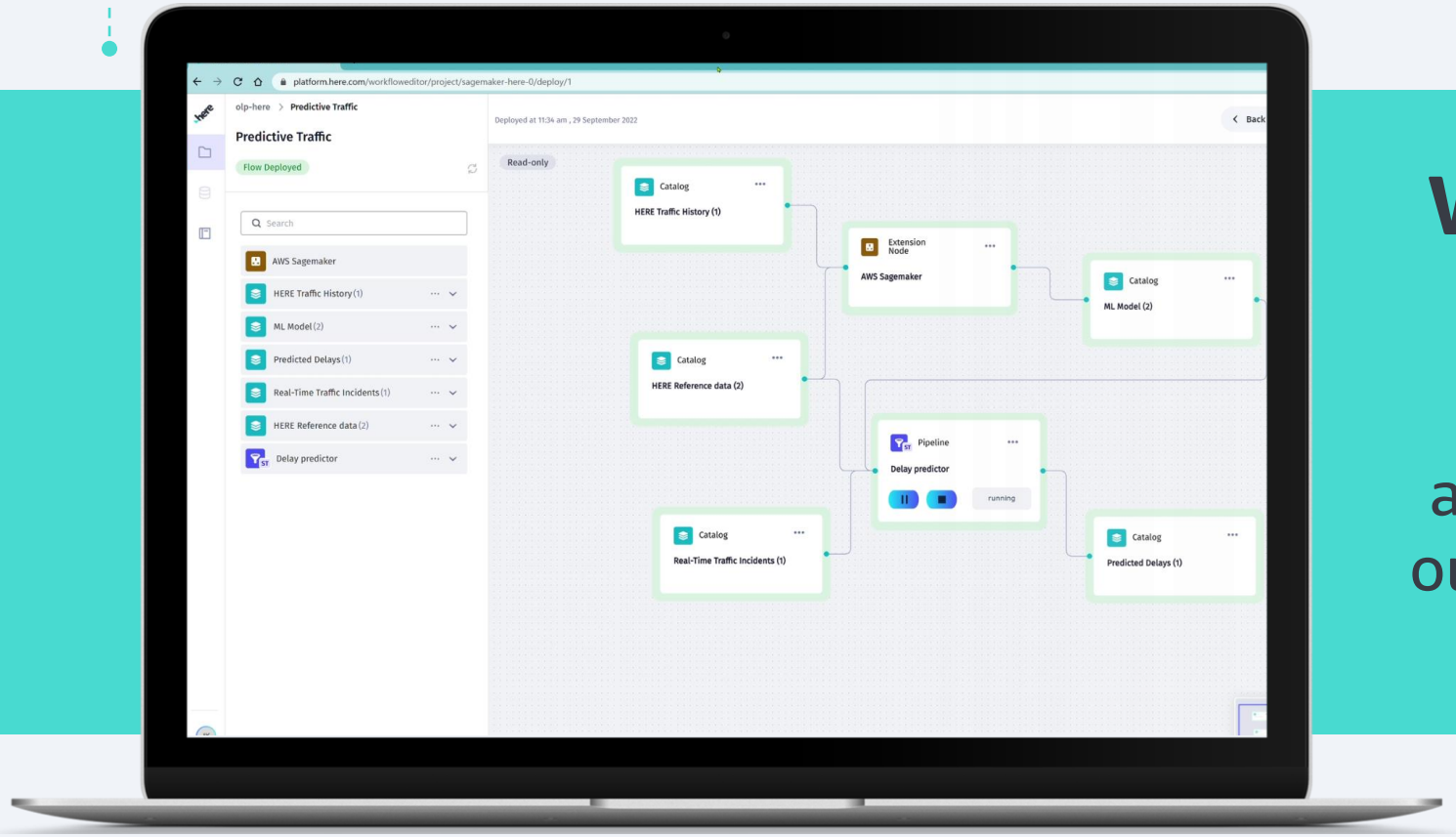
capability to aid traffic prediction

Amazon SageMaker lets us

Shadow test the predictive models

we use to power services that keep people safe on the road

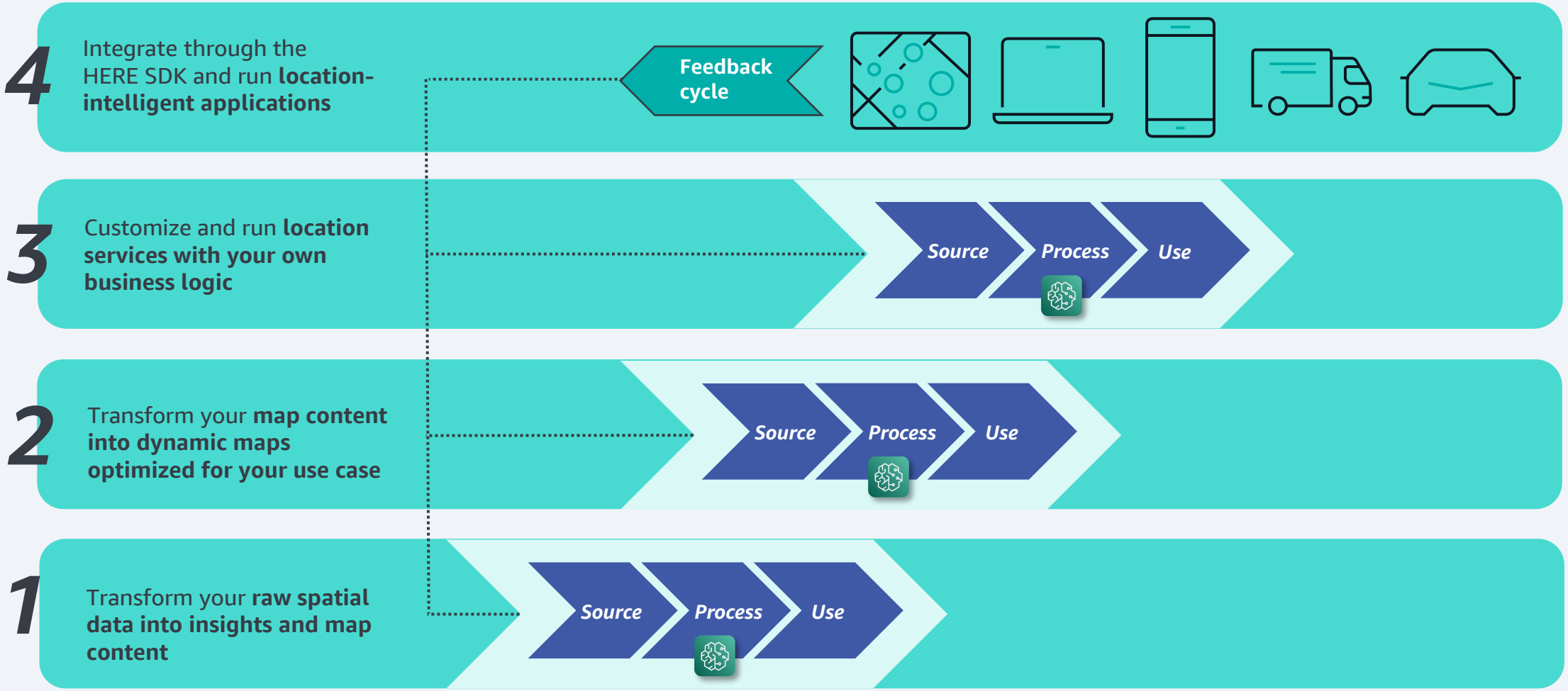




We integrated Amazon SageMaker into HERE Workspace,

a development environment on our platform for building custom maps and location services

Amazon SageMaker integrated at every step of solution creation



Predictive Traffic

Flow Deployed

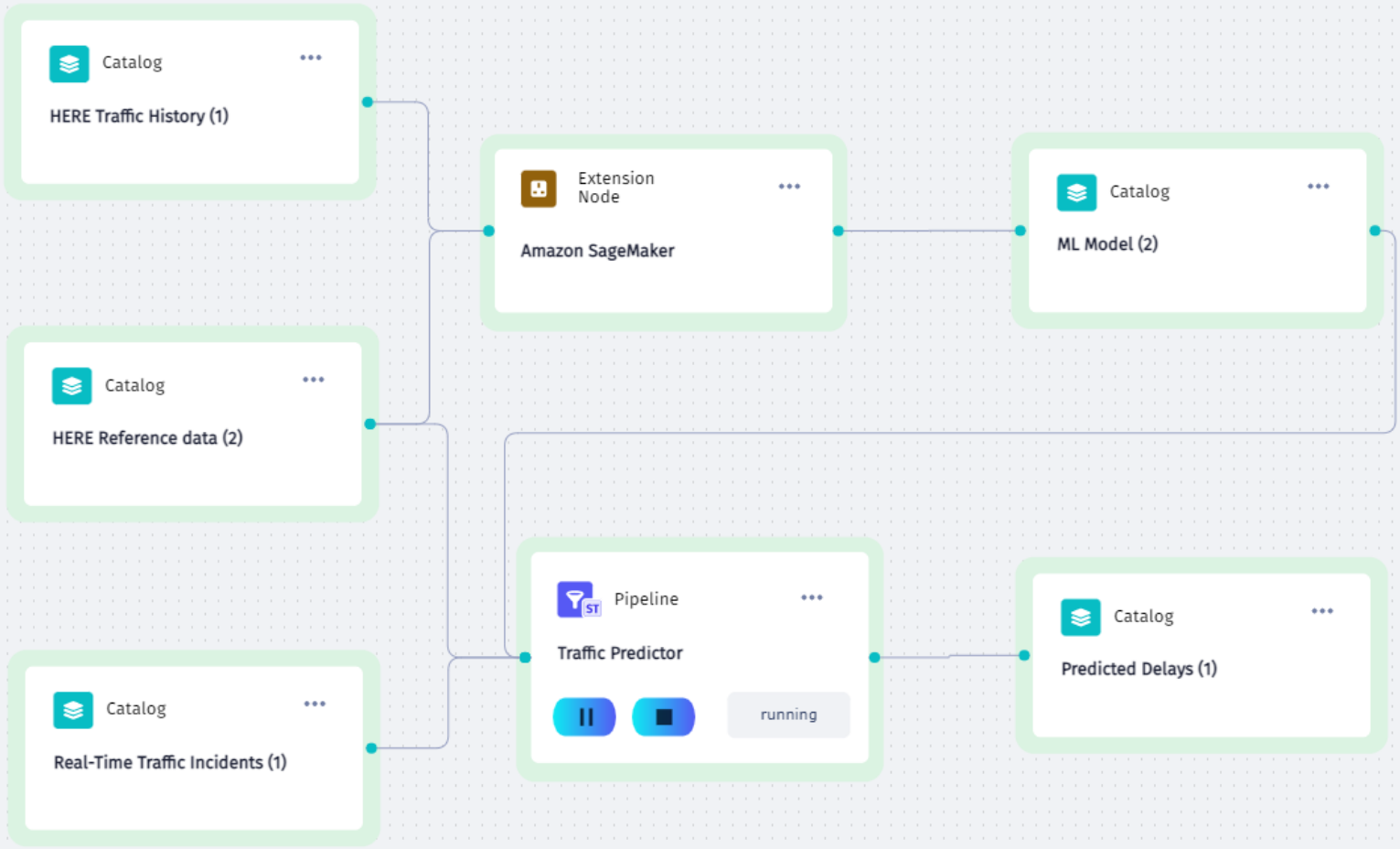
Deployed at 8:04 pm , 22 November 2022

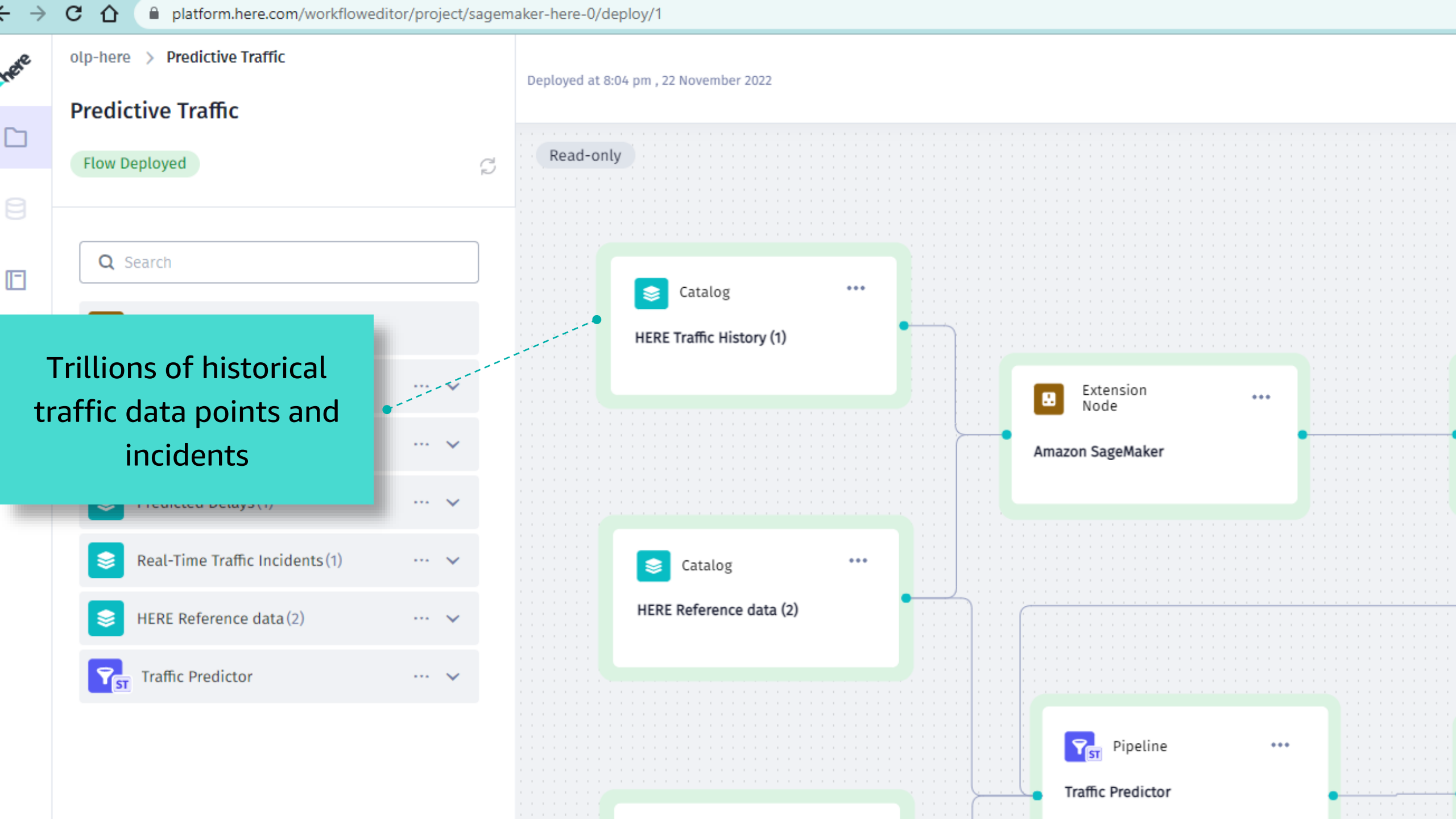
Back to flow | Undeploy

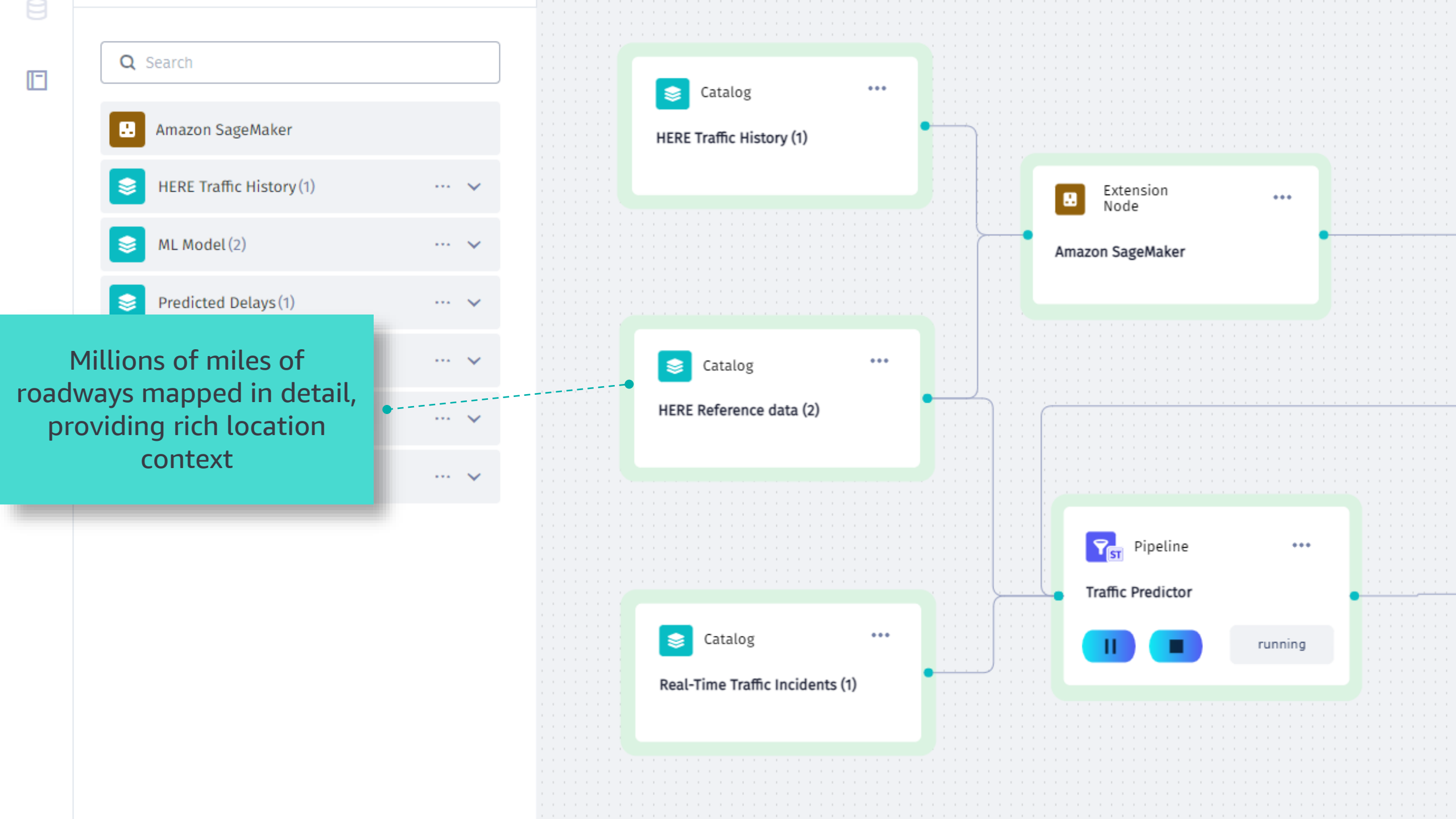
Search

- Amazon SageMaker
- HERE Traffic History (1)
- ML Model (2)
- Predicted Delays (1)
- Real-Time Traffic Incidents (1)
- HERE Reference data (2)
- Traffic Predictor

Read-only




platform.here.com/workfloweditor/project/sagemaker-here-0/deploy/1
olp-here
Predictive Traffic
Predictive Traffic
Flow Deployed
Deployed at 8:04 pm , 22 November 2022
Read-only
Search
Trillions of historical traffic data points and incidents
Real-Time Traffic Incidents (1)
HERE Reference data (2)
Traffic Predictor
Catalog
HERE Traffic History (1)
Extension Node
Amazon SageMaker
Catalog
HERE Reference data (2)
Pipeline
Traffic Predictor


Search
Amazon SageMaker
HERE Traffic History (1)
ML Model (2)
Predicted Delays (1)
Millions of miles of roadways mapped in detail, providing rich location context
Catalog
HERE Traffic History (1)
Catalog
HERE Reference data (2)
Catalog
Real-Time Traffic Incidents (1)
Extension Node
Amazon SageMaker
Pipeline
Traffic Predictor
running

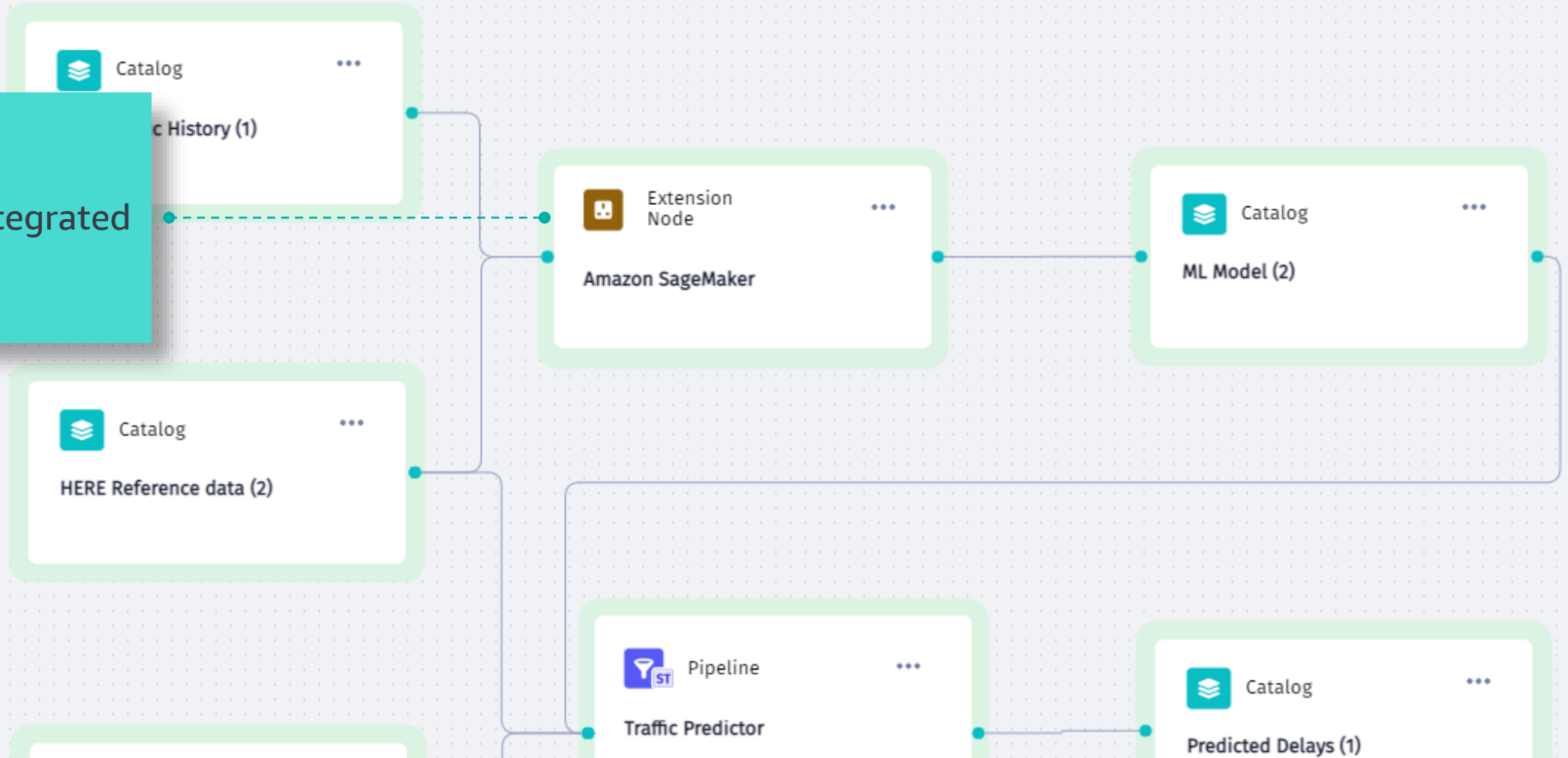
kfloweditor/project/sagemaker-here-0/deploy/1

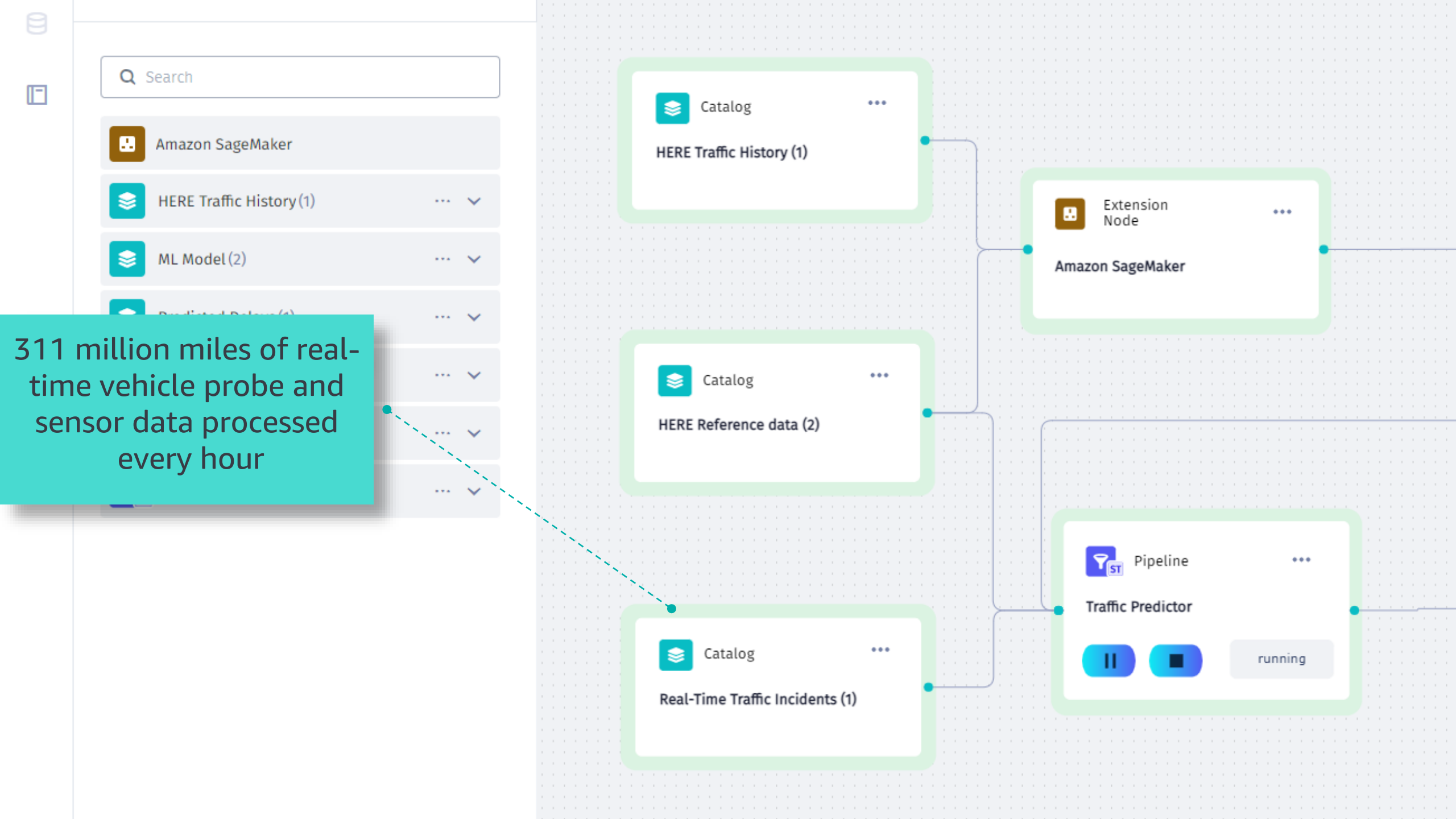
Deployed at 8:04 pm , 22 November 2022

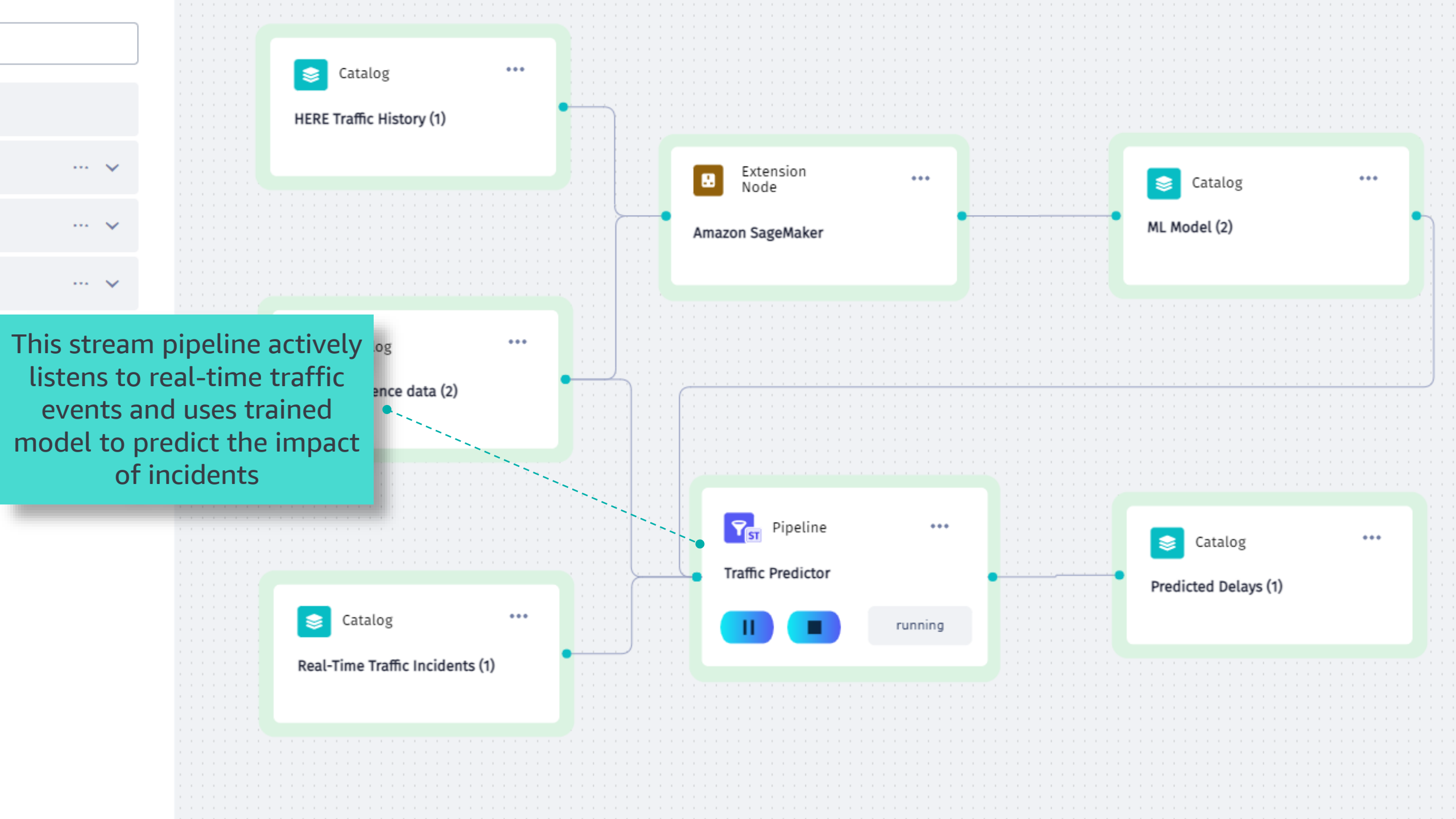
Back to flow

Read-only

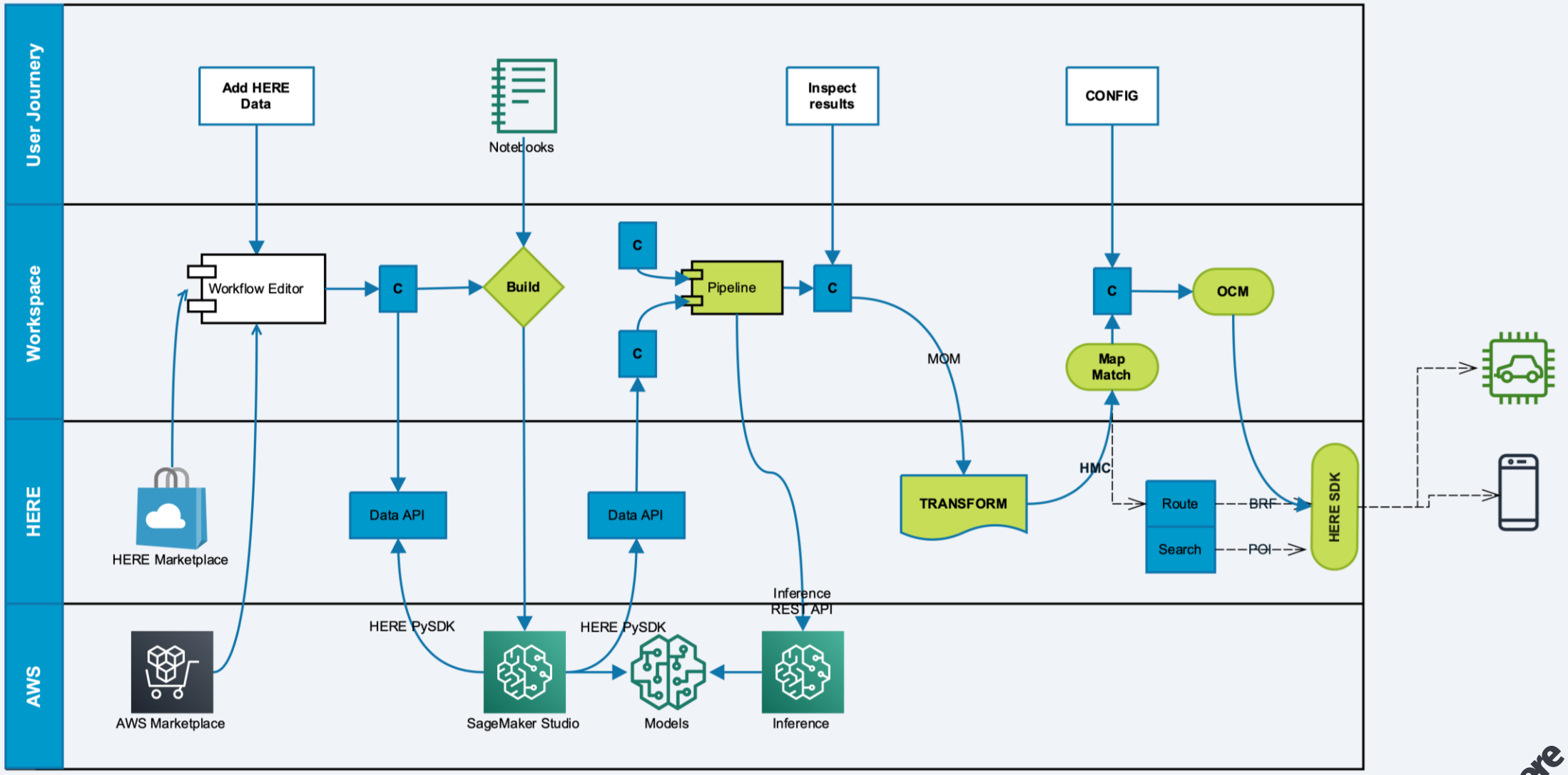
SageMaker node integrated




Search
Amazon SageMaker
HERE Traffic History (1)
ML Model (2)
311 million miles of real-time vehicle probe and sensor data processed every hour
Catalog
HERE Traffic History (1)
Extension Node
Amazon SageMaker
Catalog
HERE Reference data (2)
Catalog
Real-Time Traffic Incidents (1)
Pipeline
Traffic Predictor
running

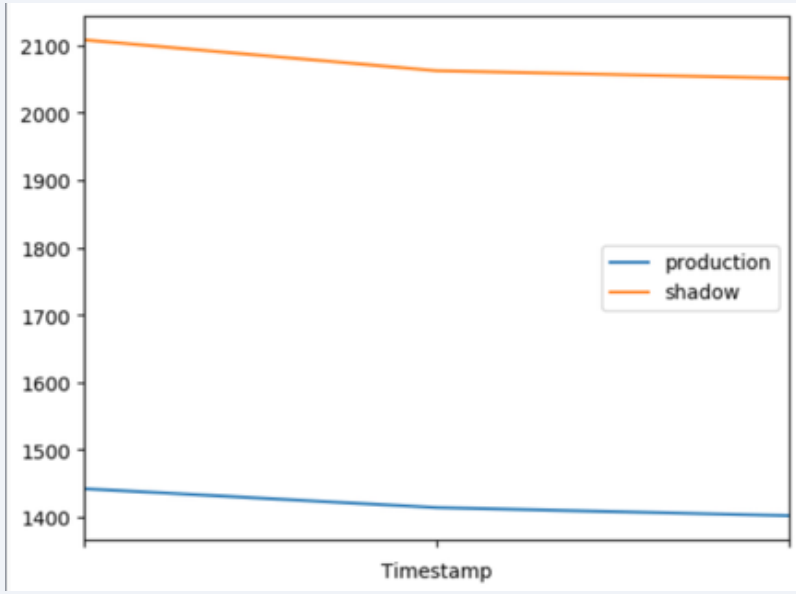

Catalog
HERE Traffic History (1)
Extension Node
Amazon SageMaker
Catalog
ML Model (2)
ence data (2)
Catalog
Real-Time Traffic Incidents (1)
Pipeline
Traffic Predictor
running
Catalog
Predicted Delays (1)
This stream pipeline actively listens to real-time traffic events and uses trained model to predict the impact of incidents

Amazon SageMaker integration architecture





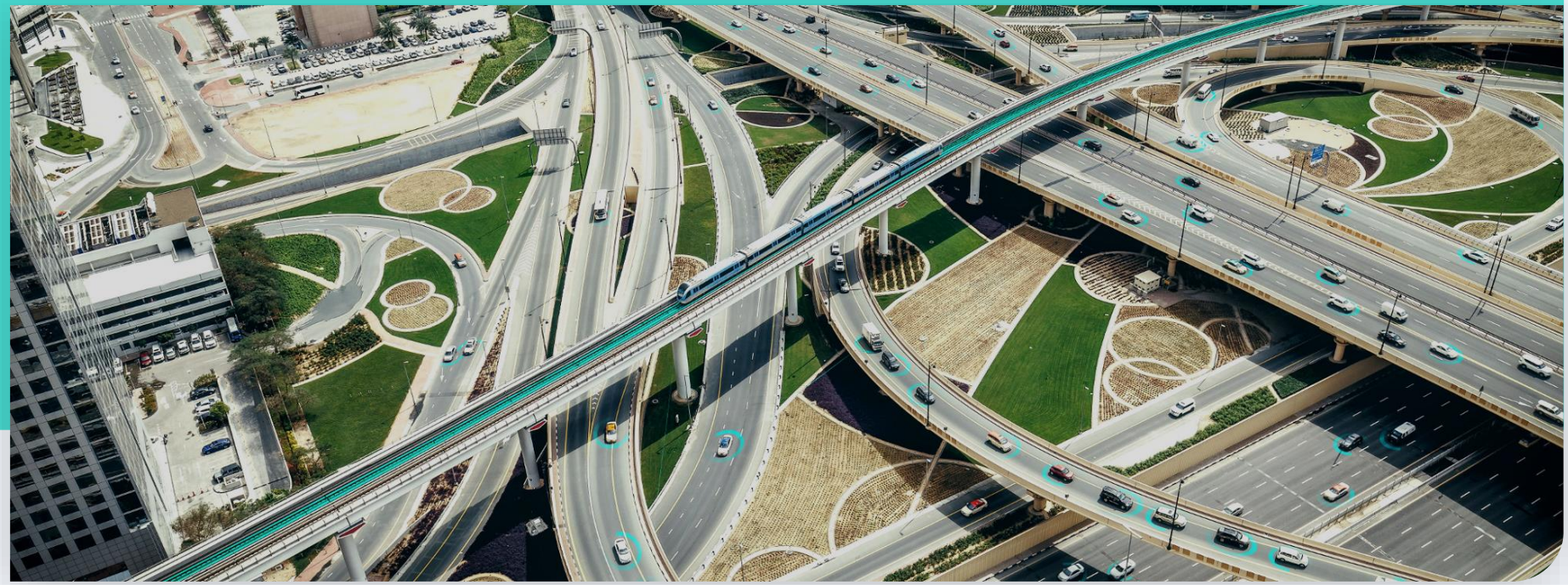
- The model predicts traffic delay in minutes
- Shadow testing allows configuration and smooth transfer of production data to both production and shadow models
- Once tested and validated for latency, alternative model can be deployed without impacting production line



The above plot shows the production and shadow model latency in **microseconds over time**; it can be inferred that more work is required in shadow model before replacing the production model

Location Intelligence + ML

- Create custom solutions
- Generate the best outcomes for your business, leveraging rich view of reality



Thank you!



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.